This listing of claims will replace all prior versions, and listings, of claims in the application:

Claims 1-16 (canceled)

Claim 17 (currently amended): A system for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

a ripper adapted to assemble ~~assembling~~ a list of word sets from unstructured content, ~~each word set~~ at least one of the word sets comprising a word and at least two ~~one~~ ~~variation on~~ non-standard capitalization variations for the word; and

an aggregator adapted to aggregate ~~aggregating~~ each word set, the aggregator including ~~comprising:~~

an analyzer adapted to identify ~~identifying at least one word set comprising significant statistics~~ non-standard capitalization variations based on at least one criteria; and

a non-standard capitalization selector adapted to select ~~selecting~~ at least ~~two such~~ one of the identified non-standard capitalization variations within the identified word set ~~having a non standard capitalization~~, and adding the ~~at least two such~~ selected at least one of the identified non-standard capitalization variations to the lexicon, wherein the lexicon includes records, each record including a word, wherein the lexicon is indexed by the words included n the records, and wherein at least one of the

records includes more than one non-standard capitalization variation.

Claim 18 (currently amended): A system according to Claim 17, further comprising:

a tokenizer adapted to tokenize ~~tokenizing~~ the excerpt into the one or more words and one or more punctuation marks.

Claim 19 (original): A system according to Claim 18, wherein hyphenated words are split into a plurality of the words.

Claim 20 (canceled)

Claim 21 (currently amended): A system according to Claim 17 ~~20~~, wherein at least one of the non-standard capitalization ~~comprises the at least one variation occurring~~ variations occurs in an excerpt having fewer than half of individual letters provided in uppercase.

Claim 22 (currently amended): A system according to Claim 17, further comprising:

a normalizer adapted to normalize ~~normalizing~~ a plurality of the words extracted relative to a source of the unstructured excerpt.

Claim 23 (currently amended): A system according to Claim 17, wherein non-standard capitalization

variations that are identified based on one or more criteria comprise ~~the set comprising significant statistics comprises~~ only those non-standard capitalization variations having at least four occurrences ~~of at least one such variation within a word set~~.

Claim 24 (currently amended): A system according to Claim 17, wherein at least one of the non-standard capitalization ~~comprises the at least one variation having~~ variations has any individual letter other than the first individual letter provided in uppercase.

Claim 25 (canceled)

Claim 26 (currently amended): A system according to Claim 17, further comprising:

a validator adapted to apply ~~applying~~ implicit rules for capitalization, and skipping each of the ~~at least two~~ non-standard capitalization variations subject to at least one such implicit rule.

Claim 27 (currently amended): A system according to Claim 26, wherein the implicit rules comprise skipping each of the ~~at least two~~ non-standard capitalization variations based on position within a sentence or phrase.

Claim 28 (currently amended): A system according to Claim 26, wherein the implicit rules comprise at least one of (A) the non-standard capitalization variation being a number, (B) the non-standard capitalization variation having no vowels, and (C) the non-standard capitalization variation constituting at least one of an article, conjunction and preposition.

Claim 29 (currently amended): A system according to Claim 26, wherein the implicit rules comprise normalizing a number of occurrences for each of the non-standard capitalization ~~at least two~~ variations ~~using at least one of a normalizing function and~~ relative to a source of the non-standard capitalization ~~each of the at least two~~ variations.

Claim 30 (currently amended): A system according to Claim 26, wherein ~~the implicit rules comprise accommodating multiple forms of capitalization for each of the at least two variations by annotating each capitalization form~~ each of the word sets includes a word and at least one non-standard capitalization variation, each of the at least one non-standard capitalization variation including ~~with~~ a frequency of occurrence count ~~and skipping those of the each of the at least two variations occurring infrequently~~.

Claim 31 (original): A system according to Claim 17, further comprising:

a hash table maintaining the lexicon.

Claim 32 (currently amended): A system according to Claim 31, ~~further comprising:~~

~~at least one record specifying at least one such word as a key into~~ wherein the hash table is indexed by words ~~, and associating at least one such variation within the word set as a preferred capitalization~~.

Claims 33-50 (canceled)

Claim 51 (new): A method comprising:

a) generating a plurality of word sets from a text corpus, each of the words sets including

- a word identified from the text corpus,
- at least one capitalization variation, and
- a frequency of occurrence of each of the at least one capitalization variation; and

b) generating a lexicon using the generated plurality of word sets, wherein the lexicon includes, for each of a plurality of words, at least one capitalization variation identified using at least one criteria, wherein at least one of the words of the lexicon includes more than one capitalization variation identified using the at least one criteria.

Claim 52 (new): The method of claim 51 wherein a capitalization variation is identified using the at least

one criteria only if it occurs at least four times in the text corpus.

Claim 53 (new): The method of claim 51 further comprising:

c) accepting a word having a capitalization defining which, if any, of the characters of the word are capitalized; and

d) performing a capitalization correction function on the word using the generated lexicon.

Claim 54 (new): The method of claim 53 wherein the act of performing a capitalization correction function includes

- determining if the capitalization of the word matches a capitalization variation in the lexicon, and
- not changing the capitalization of the word if it was determined to match a capitalization variation in the lexicon.

Claim 55 (new): The method of claim 53 wherein the act of performing a capitalization correction function includes

- determining if the capitalization of the word matches a capitalization variation in the lexicon, which capitalization variation meets a frequency criteria, and
- not changing the capitalization of the word if it was determined to match a capitalization variation in the lexicon.

Claim 56 (new): Apparatus comprising:

a) means for generating a plurality of word sets from a text corpus, each of the words sets including

- a word identified from the text corpus,
- at least one capitalization variation, and
- a frequency of occurrence of each of the at least one capitalization variation; and

b) means for generating a lexicon using the generated plurality of word sets, wherein the lexicon includes, for each of a plurality of words, at least one capitalization variation identified using at least one criteria, wherein at least one of the words of the lexicon includes more than one capitalization variation identified using the at least one criteria.

Claim 57 (new): The apparatus of claim 56 wherein a capitalization variation is identified using the at least one criteria only if it occurs at least four times in the text corpus.

Claim 58 (new): The apparatus of claim 56 further comprising:

c) means for accepting a word having a capitalization defining which, if any, of the characters of the word are capitalized; and

d) means for performing a capitalization correction function on the word using the generated lexicon.

Claim 59 (new): The apparatus of claim 58 wherein the means for performing a capitalization correction function

- determine if the capitalization of the word matches a capitalization variation in the lexicon, and
- do not change the capitalization of the word if it was determined to match a capitalization variation in the lexicon.

Claim 60 (new): The apparatus of claim 58 wherein the means for performing a capitalization correction function

- determine if the capitalization of the word matches a capitalization variation in the lexicon, which capitalization variation meets a frequency criteria, and
- do not change the capitalization of the word if it was determined to match a capitalization variation in the lexicon.